C^3: Compositional Counterfactual Contrastive Learning for Video-grounded Dialogues

Hung Le
Salesforce Research Asia
hungle@salesforce.com

Nancy F. Chen
A*STAR, Institute for Infocomm Research
nfychen@i2r.a-star.edu.sg

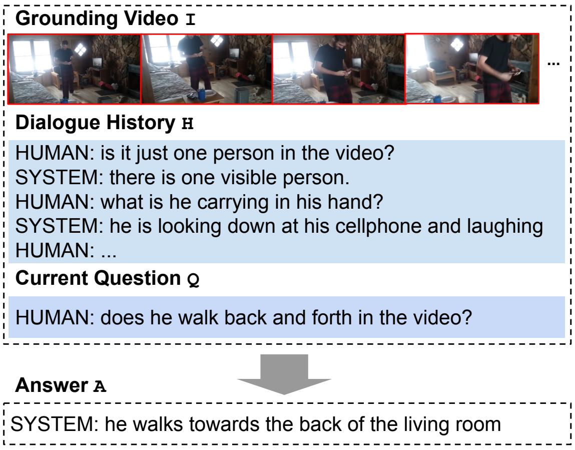
Steven C.H. Hoi
Salesforce Research Asia
shoi@salesforce.com

Abstract

Video-grounded dialogue systems aim to integrate video understanding and dialogue understanding to generate responses that are relevant to both the dialogue and video context. Most existing approaches employ deep learning models and have achieved remarkable performance, given the relatively small datasets available. However, the results are partially accomplished by exploiting biases in the datasets rather than developing multimodal reasoning, resulting in limited generalization. In this paper, we propose a novel approach of Compositional Counterfactual Contrastive Learning (C^3) to develop contrastive training between factual and counterfactual samples in video-grounded dialogues. Specifically, we design factual/counterfactual samples based on the temporal steps in videos and tokens in dialogues and propose contrastive loss functions that exploit object-level or action-level variance. Different from prior approaches, we focus on contrastive hidden state representations among compositional output tokens to optimize the representation space in a generation setting. We achieved promising performance gains on the Audio-Visual Scene-Aware Dialogues (AVSD) benchmark and showed the benefits of our approach in grounding video and dialogue context.

1 Introduction

Visual dialogue research (Das et al., 2017; Seo et al., 2017; De Vries et al., 2017; Chattopadhyay et al., 2017; Alamri et al., 2019a) aims to develop intelligent systems that can reason and answer questions about visual content in a multi-turn setting. Compared to traditional visual question answering (VQA) (Antol et al., 2015; Gao et al., 2015; Malinowski and Fritz, 2014; Zhu et al., 2016), visual dialogues bridge the gap between research and practical applications by allowing turn-based human-machine interactions. Recently, many deep learning approaches have been proposed to develop visual dialogue systems and achieved remarkable performance (Schwartz et al., 2019; Hori et al., 2019; Le et al., 2019; Li et al., 2021b). However, as these methods are heavily trained on relatively small datasets (Das et al., 2017; Alamri et al., 2019a), they are subject to inherent bias from the datasets and limited generalization into real-world applications (Zhang et al., 2016; Goyal et al., 2017). While training on large-scale data can alleviate this problem, visual dialogues are expensive to procure and require manual annotations. This challenge becomes more obvious in highly complex visual dialogue tasks such as video-grounded dialogues (Alamri et al., 2019a; Le et al., 2021) (Figure 1).



HUMAN: is it just one person in the video?
SYSTEM: there is one visible person.
HUMAN: what is he carrying in his hand?
SYSTEM: he is looking down at his cellphone and laughing
HUMAN: ...

HUMAN: does he walk back and forth in the video?

SYSTEM: he walks towards the back of the living room

Figure 1: An example of video-grounded dialogue

In recent years, we have seen increasing research efforts in contrastive learning to improve deep learning performance (Wu et al., 2018; Henaff, 2020; Chen et al., 2020; He et al., 2020). The common strategy of these methods is an objective function that pulls together representations of an anchor and "positive" samples while pushing the representations of the anchor from "negative" samples. These methods are specifically beneficial in self-supervised image representation learning. Specifically, these methods often do not require additional annotations by augmenting data of existing samples

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 548–561
September 11–15, 2023. ©2023 Association for Computational Linguistics

to create "positive" and "negative" samples. We are motivated by this line of research to improve visual dialogue systems and propose a framework of Compositional Counterfactual Contrastive Learning (C^3). C^3 includes loss functions that exploit contrastive training samples of factual and counterfactual data that are augmented to be object-variant or action-variant.

Compared to traditional deep learning tasks, a major challenge of applying contrastive learning (Wu et al., 2018; Henaff, 2020; Chen et al., 2020; He et al., 2020) in video-grounded dialogues lies in the complexity of the task. Specifically, in a discrimination task of image classification, given an image, positive samples are created based on non-adversarial transformations on this image e.g. by cropping inessential parts without changing the labels, and negative samples are randomly sampled from other image instances. However, such transformations are not straightforward to apply on visual dialogues, each of which consists of a video of spatio-temporal dimensions, a dialogue of multiple turns, and an output label in the form of natural language at the sentence level. In visual dialogues, the random sampling method, in which negative samples are created by swapping the input video and/or dialogue context with random components from other training samples, becomes too naive. In domains with high data variance like dialogues or videos, a system can easily discriminate between such positive and negative instances derived using previous approaches.

To mitigate the limitations of conventional contrastive learning in video-grounded dialogues, we propose a principled approach to generate and control negative and positive pairs by incorporating compositionality and causality (an overview of our approach can be seen in Figure 2 and 3). Specifically, we develop a structural causal model for visual dialogues by decomposing model components by object and action-based aspects. We then create hard negative samples of grounding videos by masking temporal steps that are relevant to actions mentioned in target output responses. Hard negative dialogue samples are created by masking tokens that are referenced to the entity mentioned in target output responses. Positive samples of videos and dialogues are developed similarly by masking irrelevant temporal steps or tokens for them to remain factual. Finally, based on an object or action-based variance between factual and counterfactual pairs, we only select specific hidden state representations of the target dialogue response sequence, to apply contrastive loss functions. Compared to existing approaches, our method has better control of data contrast at the granularity of object and action variance. We conducted experiments with comprehensive ablation analysis using the Audio-Visual Scene-Aware Dialogues (AVSD) benchmark (Alamri et al., 2019a) and showed that our method can achieve promising performance gains.

2 Related Work

Counterfactual Reasoning. Related to our work is the research of counterfactual reasoning. One line of research focuses on generating plausible counterfactual data to facilitate model training or evaluation. (Zmigrod et al., 2019; Garg et al., 2019; Vig et al., 2020) introduced data augmentation methods that convert gender-inflected sentences or remove identity-based tokens from sentences. The augmented data is used to study model stereotyping and improve fairness in model outputs. (Kaushik et al., 2020) crowd-sourced human annotations to minimally revise documents such that their sentiment labels are flipped. (Zeng et al., 2020; Wang and Culotta, 2020; Madaan et al., 2020) introduced data augmentation to improve model robustness in entity recognition and text classification tasks.

More related to our work are counterfactual augmentation methods in generative tasks. (Qin et al., 2019) introduced a new benchmark for counterfactual story rewriting. (Li et al., 2021a) explored augmented counterfactual dialogue goals to evaluate dialogue state tracking models. (Baradel et al., 2020) proposed a synthetic 3D environment for learning the physical dynamics of objects in counterfactual scenarios. Different from prior tasks, in the task of video-grounded dialogue, a target response is not easy to be flipped/negated, and hence, supervised learning is not straightforward. We propose to automatically develop counterfactual and factual samples and improve representation learning via unsupervised learning.

Contrastive Learning. Our work is related to the research of contrastive learning in deep learning models. The research is particularly popular in self-supervised learning of image representations (Wu et al., 2018; Hjelm et al., 2019; Henaff, 2020; Chen et al., 2020; He et al., 2020; Khosla et al., 2020). These methods do not require additional annotations but aim to improve representations

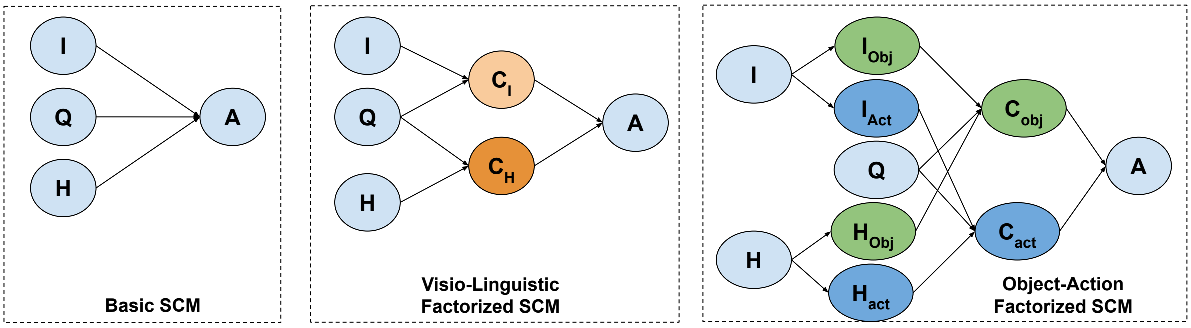


Figure 2: **SCMs of video-grounded dialogues:** Left: Basic SCM without factorization. Middle: SCM factorized by visual and textual context. Right: SCM factorized by object and action-level information. I: video input, Q: question input, H: dialogue history, C: contextualized information, and A: target response. For simplicity, we do not demonstrate independent noise variables U and the subscript t.

through loss functions. The loss functions are often inspired by noise contrastive estimation (NCE) (Gutmann and Hyvärinen, 2010) and applied in lower-dimensional representation space. In the language domain, similar loss functions have been introduced to improve word embeddings (Mnih and Kavukcuoglu, 2013) and sentence embeddings (Logeswaran and Lee, 2018). More related to our work is (Huang et al., 2018; Liu and Sun, 2015; Yang et al., 2019; Lee et al., 2021), introducing positive and negative pairs of sentences for contrastive learning in generative tasks such as language modelling, word alignment, and machine translation. In the multimodal research domains, our work is related to contrastive learning methods introduced by (Zhang et al., 2020; Gokhale et al., 2020; Liang et al., 2020; Gupta et al., 2020). Specifically, our work complements (Zhang et al., 2020) by incorporating causality into contrastive learning. However, we focus on a very different task of video-grounded dialogues that involves turn-based question-answering. The task requires multimodal reasoning performed on both dialogue context and video context. Moreover, we improve models by tightly controlling data variance by adopting compositionality and our loss functions optimize hidden state representations of decoding tokens by their object or action-based semantics.

3 Method

3.1 Problem Definition

In a video-grounded dialogue task (Alamri et al., 2019a; Le et al., 2021), the inputs consist of a dialogue $\mathcal{D}$ and the visual input of a video $\mathcal{I}$. Each dialogue contains a sequence of dialogue turns, each of which is a pair of question $\mathcal{Q}$ and answer $\mathcal{A}$. At each dialogue turn t, we denote the dialogue context $\mathcal{H}_t$ as all previous dialogue turns $\mathcal{H}_t = \{(\mathcal{Q}_i, \mathcal{A}_i)\}|_{i=1}^{i=t-1}$. The output is the answer $\hat{\mathcal{A}}_t$ to answer the question of the current turn $\mathcal{Q}_t$. The objective of the task is the generation objective that output answers of the current turn:

$$\hat{\mathcal{A}}_t = \arg\max_{\mathcal{A}_t} P(\mathcal{A}_t|\mathcal{I}, \mathcal{H}_t, \mathcal{Q}_t; \theta) \qquad (1)$$

3.2 Structural Causal Model

We first cast a visual dialogue model as a structural causal model (SCM) (Pearl, 2009) to explore the potential factors that affect the generation of target dialogue responses in a dialogue system. By definition, an SCM consists of random variables $V = \{V_1, ..., V_N\}$ and corresponding independent noise variables $U = \{U_1, ..., U_N\}$. We assume an SCM of a directed acyclic graph (DAG) structure. In this structure, causal functions are defined as $F = \{f_1, ..., f_N\}$ such that $V_i = f_i(P_i, U_i)$ where $P_i = \{V_p\} \subset V$ are the parent nodes of V_i in the DAG. Using this definition of SCM, we develop three SCM structures for a video-grounded dialogue system in Figure 2.

The *Basic SCM* is directly derived from the objective function (1). The *VL-SCM* adopts a question-aware reasoning process that partitions visual and language reasoning based on question information as the common cause. A limitation of VL-SCM is that it does not account for the interactions of components such as object and action abstracts that are embedded in visual context C_I and linguistic context C_H. This drawback becomes more significant in scenarios in which question information is highly dependent on prior turns in the dialogue history. Specifically, in questions that involve references, including object refer-

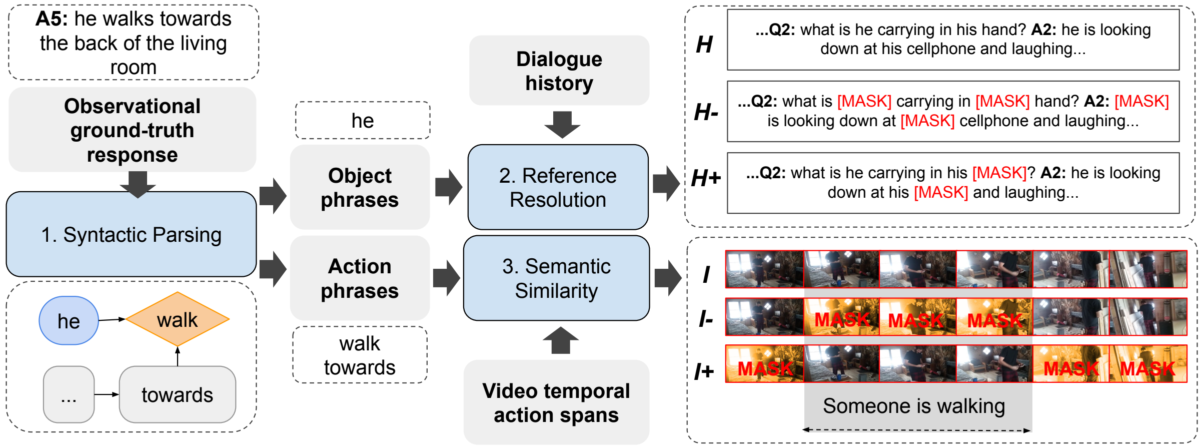

A5: he walks towards the back of the living room
Observational ground-truth response
1. Syntactic Parsing
he
walk
...
towards
he
Object phrases
Action phrases
walk towards
Dialogue history
2. Reference Resolution
3. Semantic Similarity
Video temporal action spans
H
...Q2: what is he carrying in his hand? A2: he is looking down at his cellphone and laughing...
H-
...Q2: what is [MASK] carrying in [MASK] hand? A2: [MASK] is looking down at [MASK] cellphone and laughing...
H+
...Q2: what is he carrying in his [MASK]? A2: he is looking down at his [MASK] and laughing...
I
I-
MASK
MASK
MASK
I+
MASK
MASK
MASK
Someone is walking

Figure 3: **Counterfactual generation:** An overview of our factual and counterfactual dialogue/video generation.

ences ("does *she* interact with *the woman in red*?") and action references ("what does the boy do after *that*?"), VL-SCM is not optimal to integrate dialogue and video context to solve component references such as "she" and "that". To address this drawback, we propose an *OA-SCM* that is factorized by object-action contextual information (Figure 2, right). The causal functions f_H^{obj} and f_H^{act} can be a simple text parser that map tokens into object-based tokens or action-based tokens s.t. $\mathcal{H}_{obj} = f_H^{obj}(\mathcal{H})$ and $\mathcal{H}_{act} = f_H^{act}(\mathcal{H})$. Similarly, f_I^{obj} and f_I^{act} are causal functions that map bounding boxes or temporal steps into object-based or action-based contents. In Section 3.3, we show that OA-SCM structure provides a framework to develop *partially counterfactual* training samples.

3.3 Counterfactual Augmentation

An overview of our augmentation process can be seen in Figure 3.

Decomposing observational target response. First, at each dialogue turn t, the ground-truth dialogue response $\mathcal{A}_t$ are passed to a syntactic parser such as the Stanford parser system [1]. The output includes grammatical components, such as subjects, verbs, and modifiers, in the form of a dependency tree. We prune the dependency tree to remove inessential parts and extract a set of object phrases $\mathcal{A}_{t,obj}$, and action-based phrases $\mathcal{A}_{t,act}$.

Generating counterfactual dialogue. Based on $\mathcal{A}_{t,obj}$, we apply a pretrained reference resolution model e.g. (Clark and Manning, 2016), to the dialogue context $\mathcal{H}_t$ to identify any references from past dialogue turns to any objects in $\mathcal{A}_{t,obj}$. For instance, in Figure 2, the object "he" identified in $\mathcal{A}_t$ are mapped to different token positions in prior dialogue turns, e.g. "his" in the text span "his hand" in the second question turn. All referenced tokens in dialogue context $\mathcal{H}_t$ are replaced by a MASK vector and the resulting dialogue context is denoted as counterfactual sample $\mathcal{H}_t^-$. We also used the pretrained reference resolution model to select any object tokens in $\mathcal{H}_t$ that are not mapped to $\mathcal{A}_{t,obj}$. These objects are considered irrelevant to $\mathcal{A}_t$ and they are replaced by the MASK vector from $\mathcal{H}_t$ and the resulting dialogue is denoted as a factual sample $\mathcal{H}_t^+$.

Generating counterfactual video. To create a counterfactual video sample, we first identify the temporal steps from the video that are semantically relevant to action phrases in $\mathcal{A}_{t,act}$. We obtain the annotation of temporal action spans from video, which can be retrieved from a pretrained temporal localization model (Shou et al., 2016) or is readily available in existing video benchmarks (Sigurdsson et al., 2016). The action span annotations consist of a set of action labels $Y_{i,act}$, each of which is mapped to a start and end time (t_i^s, t_i^e). Temporal segments that are deemed necessary to generate $\mathcal{A}_t$ is the union of all time spans from the set $S = \{(t_j^s, t_j^e)\}$ for all $Y_{j,act}$ that is semantically similar to $\mathcal{A}_{t,act}$. To identify similar pairs, we adopted cosine similarity scores between pretrained Glove embedding vectors of $Y_{j,act}$ and $\mathcal{A}_{t,act}$. During video feature encoding, any features of temporal steps sampled within S are replaced with a MASK vector, and resulting video features are noted as encoded features of counterfactual video $\mathcal{I}^-$. Factual video $\mathcal{I}^+$ are created similarly but for video parts irrelevant to $\mathcal{A}_t$, that is $I \setminus S$.

[1] https://nlp.stanford.edu/software/lex-parser.shtml

By the definition of OA-SCM from Section 3.2, we can denote $\mathcal{H}_t^- = \mathcal{H}_{t,obj}^- + \mathcal{H}_{t,act}$ and $\mathcal{H}_t^+ = \mathcal{H}_{t,obj}^+ + \mathcal{H}_{t,act}$; and $\mathcal{I}^- = \mathcal{I}_{obj} + \mathcal{I}_{act}^-$ and $\mathcal{I}^+ = \mathcal{I}_{obj} + \mathcal{I}_{act}^+$. Note that we follow (Hsieh et al., 2018) and assume object information such as object appearance and shape are typically embedded in any video frame. In this case, $\mathcal{I}_{obj}$ is unchanged and can be obtained from either $I \setminus S$ or S. In Section 3.4, we show that these partially counterfactual formulations enable a compositional contrastive learning approach.

3.4 Contrastive Learning

In this section, we introduce a contrastive learning method that exploits the compositional hidden states between factual and counterfactual samples. We extend the objective function (1) to express the auto-regressive decoding process:

$$\begin{aligned}\hat{\mathcal{A}}_t &= \arg\max_{\mathcal{A}_t} P(\mathcal{A}_t|\mathcal{I}, \mathcal{H}_t, \mathcal{Q}_t; \theta) \\ &= \arg\max_{\mathcal{A}_t} \prod_{m=1}^{L_{\mathcal{A}}} P_m(w_m|\mathcal{A}_{t,<m}, \mathcal{I}, \mathcal{H}_t, \mathcal{Q}_t; \theta)\end{aligned}$$

Each target response $\mathcal{A}$ is represented as a sequence of token or word indices $\{w_m\}|_{m=1}^{m=L} \in |\mathbb{V}|$, where L is the sequence length and $\mathbb{V}$ is the vocabulary set. The conditional probability P_m is defined as:

$$P_m = \text{softmax}(Wk_m + b) \in \mathbb{R}^{|\mathbb{V}|} \tag{2}$$

$$k_m = \theta_{\text{decode}}(w_{m-1}, \theta_{\text{encode}}(\mathcal{I}, \mathcal{H}_t, \mathcal{Q}_t)) \tag{3}$$

where k_m is the hidden state at decoding position m and d is the embedding dimension of the hidden state. In this generative setting, we then explain 2 different ways of contrastive learning:

Sentence-level contrast. This approach learns the representations of the hidden states by contrasting a linear transformation of an aggregated vector of hidden states following an NCE framework:

$$\mathcal{L}_{\text{nce}}^{\text{sent}} = -\log \frac{e^{\text{sim}(z,z^+)}}{e^{\text{sim}(z,z^+)} + e^{\text{sim}(z,z^-)}} \tag{4}$$

where $\text{sim}(,)$ is the cosine similarity score and z is the output of an aggregation function Agg: $z = \text{Agg}(U)$ where $U \in \mathbb{R}^{d_{\text{nce}} \times L_{\mathcal{A}}}$ and $u_m = \text{MLP}_{\text{nce}}(k_m) \in \mathbb{R}^{d_{\text{nce}}}$. z^+ and z^- are obtained similarly by passing k_m^+ and k_m^- to the same MLP and aggregation function. k_m^+ and k_m^- are obtained by passing factual and counterfactual video pairs into (3): $k_m^+ = \theta_{\text{decode}}(w_{m-1}, \theta_{\text{encode}}(\mathcal{I}^+, \mathcal{H}_t, \mathcal{Q}_t))$ and $k_m^- = \theta_{\text{decode}}(w_{m-1}, \theta_{\text{encode}}(\mathcal{I}^-, \mathcal{H}_t, \mathcal{Q}_t))$. In cases of augmentation with factual and counterfactual dialogues, we obtain k_m^+ and k_m^- by replacing $\mathcal{H}$ with $\mathcal{H}^+$ and $\mathcal{H}^-$ in (3). Agg is an aggregation function that collapses hidden states into a single vector, e.g. average pooling (Lee et al., 2021; Zhang et al., 2020). We follow (Khosla et al., 2020) to normalize z, z^+, z^- to lie on the unit hypersphere. To reflect this contrastive learning approach against the VL-SCM, we can assume $\mathcal{C} \cong K$ and (4) essentially exploits the contrast between $\mathcal{C}^+$ and $\mathcal{C}^-$.

Compositional contrast. We note that the above approach does not consider compositionality in the target output response $\mathcal{A}$. Since we are using the same observational output w_{m-1} to obtain k_m, k_m^-, and k_m^-, we can remove the Agg function and apply a token-level pairwise contrastive loss between pairs of $(z_m = u_m, z_m^+ = u_m^+)$ and $(z_m = u_m, z_m^- = u_m^-)$. In this strategy, we formulate a loss function for action variance between $\mathcal{I}^+$ and $\mathcal{I}^-$, and one for object variance between $\mathcal{H}^+$ and $\mathcal{H}^-$:

$$\mathcal{L}_{\text{nce}}^{\text{act}} = -\frac{1}{|D_{act}|} \sum_{i \in D_{act}} \log \frac{e^{\text{sim}(z_i,z_i^+)}}{e^{\text{sim}(z_i,z_i^+)} + e^{\text{sim}(z_i,z_i^-)}}$$

$$D_{act} = \{\text{idx}(w_i) : w_{i-1} \in \mathcal{A}_{t,act}\} \tag{5}$$

$$\mathcal{L}_{\text{nce}}^{\text{obj}} = -\frac{1}{|D_{obj}|} \sum_{j \in D_{obj}} \log \frac{e^{\text{sim}(z_j,z_j^+)}}{e^{\text{sim}(z_j,z_j^+)} + e^{\text{sim}(z_j,z_j^-)}}$$

$$D_{obj} = \{\text{idx}(w_j) : w_{j-1} \in \mathcal{A}_{t,obj}\} \tag{6}$$

where $idx(w_m)$ returns the index of w_m in $\mathcal{A}_t$. Note that in (5) and (6), we adopt a hypothetical strategy by obtaining hidden states given *input* tokens are either in $\mathcal{A}_{t,act}$ or $\mathcal{A}_{t,obj}$. An alternative approach is to consider hidden states that are expected to produce *prospective* tokens $w_m \in \mathcal{A}_{t,act}/\mathcal{A}_{t,obj}$, i.e. $D'_{act} = \{\text{index}(w_i) : w_i \in \mathcal{A}_{t,act}\}$ and $D'_{obj} = \{\text{index}(w_j) : w_j \in \mathcal{A}_{t,obj}\}$. We conducted experiments with both strategies and explained our findings in the next section. Note that we can connect the compositional contrastive learning approach against the OA-SCM (Section 3.2) by denoting $\mathcal{C}_{act} \cong \{k_i\} \forall i \in D_{act}$ and $\mathcal{C}_{obj} \cong \{k_j\} \forall j \in D_{obj}$. Therefore, (5) essentially exploits the contrast between $\mathcal{C}_{act}^+$ and $\mathcal{C}_{act}^-$, and (6) for the contrast between $\mathcal{C}_{obj}^+$ and $\mathcal{C}_{obj}^-$.

4 Experiments

Dataset and Experimental Setup. We used the Audio-Visual Sene-Aware Dialogue (AVSD)

	Train	**Train$_{\text{aug}}^{\text{video}}$**	**Train$_{\text{aug}}^{\text{dial}}$**	**Val**	**Val$_{\text{aug}}^{\text{video}}$**	**Val$_{\text{aug}}^{\text{dial}}$**	**Test**
#Dialogs	7,659	7,145	6,411	1,787	1,709	1,557	1,710
#$(\mathcal{I}, \mathcal{H}_t, \mathcal{Q}_t, \mathcal{A}_t)$	76,590	28,163	18,397	17,870	7,383	4,912	6,745

Table 1: Summary of the AVSD benchmark with augmented counterfactual video/dialogue data

dataset (Alamri et al., 2019b) to benchmark video-grounded dialogue systems. The dataset contains 10-turn dialogues, each of which is grounded on one video from the Charades dataset (Sigurdsson et al., 2016). We used the standard visual features I3D (Carreira and Zisserman, 2017) to represent the video input. Note that compared to (Alamri et al., 2019b), we followed the setting of AVSD in the 7^{th} Dialogue System Technology Challenge (DSTC7) (Yoshino et al., 2019), which requires generating a response rather than selecting from a candidate set. We also did not use video caption as an input as the caption is typically not easy to obtain in applications. A summary of the dataset can be seen in Table 1.

All model parameters, except the visual feature extractor of a pretrained I3D model, are initialized with uniform distribution (Glorot and Bengio, 2010). Our approach can be applied to different model architectures, as long as the hidden states of individual decoding tokens are available for contrastive learning. We used MTN (Le et al., 2019), which is a Transformer adaptation of the traditional RNN-based dialogue systems, as our base model. Finally, we evaluated models with objective metrics, including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015). These metrics are found to correlate well with human judgment (Alamri et al., 2019b).

Creating Counterfactual Data. We created counterfactual data for the training split and validation split of the AVSD benchmark. Specifically, from the original data, we identified invalid samples that are not sufficient for factual and counterfactual transformations. Examples of invalid samples are ones with ambiguous actions in target responses (e.g. "I am not sure what he is doing"), or ones without object references to prior turns (e.g. "there is only a single person in the video"). These samples are discarded and the remaining data is processed as described in Section 3.3. The overall statistics of augmented train and validation splits can be seen in Table 1. Note that the number of samples with augmented videos and dialogues are different as some samples contain valid actions but no object references (e.g. "the man is walking around the kitchen"), and vice versa.

Evaluating with Counterfactual Data. First, using augmented data, we evaluated models trained only with the original data. Motivated by (Kaushik et al., 2020; Vig et al., 2020; Agarwal et al., 2020), we designed this set of experiments to gauge the model performance under adversarial (counterfactual) samples and favourable (factual) samples and to observe the effects of our transformation methods. Specifically, we trained an MTN model (Le et al., 2019) on the original training data and evaluate the model on an augmented validation set. To fairly compare the results, we create a shared validation set in which each sample is augmented with both video and dialogue factual and counterfactual pairs. Essentially, this set is the intersection $\text{Val}_{\text{aug}}^{\text{v+d}} = \text{Val}_{\text{aug}}^{\text{video}} \cap \text{Val}_{\text{aug}}^{\text{dial}}$. Using the CIDEr metric (Vedantam et al., 2015), We noted the MTN model pretrained on original training data achieves 0.996 and 1.086 score in the original test and validation set respectively. However, as noted from Table 2, the performance drops to 0.779 when evaluating on the validation set $\text{Val}_{\text{aug}}^{\text{v+d}}$ even with the original video-dialogue pair $(\mathcal{I}, \mathcal{H})$. This performance drop indicates that the subset contains more challenging instances that require reasoning in dialogues and videos.

The performance decreases to 0.760 when tested with $\mathcal{I}^-$ and increases to 0.782 when tested with $\mathcal{I}^+$, keeping the $\mathcal{H}$ unchanged. When tested with videos that are masked at random temporal steps $\mathcal{I}_{\text{rand}}^-$, the result only reduces to 0.773, less than $\mathcal{I}^-$. This illustrates higher counterfactual impacts in $\mathcal{I}^-$ than in $\mathcal{I}_{\text{rand}}^-$. We also observed that model performance with counterfactual videos $\mathcal{I}^-$ is higher than cases with no video at all, $\mathcal{I}^0$. This observation demonstrates the factorization formulation of our SCM in which $\mathcal{I}^-$ is partially counterfactual, containing useful information, i.e. $\mathcal{I}_{obj}$, than $\mathcal{I}^0$, to support response generation.

When tested with dialogue transformations, we have similar observations with $\mathcal{H}^-$, $\mathcal{H}^+$, $\mathcal{H}_{\text{rand}}^-$, and $\mathcal{H}^0$. Specifically, following our SCM structure,

Video augmentation + original dialogue					Video augmentation + no dialogue				
$(\mathcal{I}, \mathcal{H})$	$(\mathcal{I}^-, \mathcal{H})$	$(\mathcal{I}^0, \mathcal{H})$	$(\mathcal{I}^+, \mathcal{H})$	$(\mathcal{I}^-_{\text{rand}}, \mathcal{H})$	$(\mathcal{I}, \mathcal{H}^0)$	$(\mathcal{I}^-, \mathcal{H}^0)$	$(\mathcal{I}^0, \mathcal{H}^0)$	$(\mathcal{I}^+, \mathcal{H}^0)$	$(\mathcal{I}^-_{\text{rand}}, \mathcal{H}^0)$
0.779	0.760	0.733	0.782	0.773	0.724	0.708	0.693	0.722	0.710
Dialogue augmentation + original video					Dialogue augmentation + no video				
$(\mathcal{I}, \mathcal{H})$	$(\mathcal{I}, \mathcal{H}^-)$	$((\mathcal{I}, \mathcal{H}^0)$	$(\mathcal{I}, \mathcal{H}^+)$	$(\mathcal{I}, \mathcal{H}^-_{\text{rand}})$	$(\mathcal{I}^0, \mathcal{H})$	$(\mathcal{I}^0, \mathcal{H}^-)$	$(\mathcal{I}^0, \mathcal{H}^0)$	$(\mathcal{I}^0, \mathcal{H}^+)$	$(\mathcal{I}^0, \mathcal{H}^-_{\text{rand}})$
0.779	0.764	0.724	0.788	0.778	0.733	0.722	0.693	0.739	0.734

Table 2: **Validation results with augmentation data:** $\mathcal{I}$: original video input, $\mathcal{I}^{-/+}$: counterfactual/factual video following Section 3.3, $\mathcal{I}^-_{\text{rand}}$: counterfactual video by masking random temporal steps, $\mathcal{I}^0$: no video input; $\mathcal{H}$: original dialogue input, $\mathcal{H}^{-/+}$: counterfactual/factual dialogue following Section 3.3, $\mathcal{H}^-_{\text{random}}$: counterfactual dialouge by masking random tokens, $\mathcal{H}^0$: no dialogue input. All results are in CIDEr score.

#	Contrast pair	Contrast loss	Hidden states	B-1	B-2	B-3	B-4	M	R	C
A	-	-	-	0.695	0.558	0.455	0.376	0.253	0.534	0.996
B	$\mathcal{I}^+, \mathcal{I}^-$	NCE	D_{act}	**0.709**	**0.577**	**0.476**	**0.398**	**0.262**	**0.549**	**1.040**
C	$\mathcal{I}^+, \mathcal{I}^-$	NCE	D'_{act}	0.697	0.565	0.462	0.381	0.254	0.538	1.003
D	$\mathcal{I}^+, \mathcal{I}^-$	NCE	D_{obj}	0.701	0.565	0.462	0.383	0.256	0.541	1.011
E	$\mathcal{I}^+, \mathcal{I}^-$	NCE	D	0.699	0.566	0.465	0.386	0.253	0.539	1.008
F	$\mathcal{I}^+, \mathcal{I}^-_{\text{rand}}$	NCE	D_{act}	0.693	0.563	0.464	0.388	0.254	0.538	1.010
G	$\mathcal{I}^+, \mathcal{I}^0$	NCE	D	0.700	0.566	0.463	0.383	0.256	0.538	1.019
H	$\mathcal{I}^+, \mathcal{I}^0_{\text{rand}}$	NCE	D	0.695	0.563	0.463	0.385	0.253	0.538	0.998
I	$\mathcal{I}^+, \mathcal{I}^-$	S-NCE	D_{act}	0.695	0.567	0.467	0.389	0.255	0.54	1.014
J	$\mathcal{I}^+, \mathcal{I}^-$	L1-PD	D_{obj}	0.705	0.569	0.465	0.385	0.258	0.543	1.005

Table 3: **Contrastive learning with counterfactual videos:** We experimented with variants of contrastive video pairs, hidden state sampling, and contrast loss. Metrics: B-n: BLEU-n, M: METEOR, R: ROUGE-L, C: CIDEr.

we show that $\mathcal{H}^-$ is partially counterfactual. To isolate the impacts of video/dialogue augmentations, we also tested models with tuples that are paired with zero dialogue context/video input ($\mathcal{H}_0/\mathcal{I}_0$). In these isolated experiments, we still observe consistent performance patterns among different variants of augmented video/dialogues, validating our factorization SCM and the effectiveness of augmentation techniques.

Contrastive Learning with Counterfactual Videos. In these experiments, we combined the task objective loss with our proposed contrastive learning approach that exploits action-based data contrast between $\mathcal{I}^+$ and $\mathcal{I}^-$. From Table 3, we have the following observations: **1)** First, when applying contrastive learning on augmented counterfactual data following our NCE function (5) (Row B), the model outperforms one which was trained with only original training data (Row A). This demonstrates the positive impacts of our C^3 learning approach through better generated target responses. **2)** When using the indices of hidden states based on prospective tokens (D'_{act}) (Row C), the performance gain decreases. This can be explained as hidden states in D'_{act} positions represent contextual information that *potentially*, but not absolutely, generate an action token. However, hidden states in D_{act} positions already assume a hypothetical input action token (w_{i-1} in (5)), and hence, a contrastive learning on these hidden states is more stable. **3)** we observed marginal performance gains when changing hidden state indices to indices of object tokens D_{obj} (Row D) or to hidden states of all tokens D (Row E). This observation verifies our factorized SCM framework as $\mathcal{I}^-$ and $\mathcal{I}^+$ are formulated to be action-variant specifically. Training them based on object variance or generic variance might lead to unstable representation learning and trivial performance gains.

4) Consistent with our observations from Table 2, contrastive learning applied to counterfactual videos with random masked temporal steps $\mathcal{I}^-_{\text{rand}}$ (Row F) results in very low performance gain. **5)** When we applied contrastive learning between $\mathcal{I}^+$ and naive counterfactual samples, including zero video input $\mathcal{I}^0$ (Row G) or video input sample from other training instance I^0_{rand} (Row H), the results only increases marginally compared to results with $\mathcal{I}^-$. **6)** We experimented with sentence-level contrast (S-NCE) (as in (4)) in which all hidden states are considered and collapsed to a single vector, as similarly used by (Lee et al., 2021; Zhang et al., 2020). We observed that this loss formulation (Row I), is not effective in our task, illustrating the benefits of using compositional representations of decoding tokens. **7)** Fi-

#	Contrast pair	Contrast loss	Hidden states	B-1	B-2	B-3	B-4	M	R	C
A	-	-	-	0.695	0.558	0.455	0.376	0.253	0.534	0.996
B	$\mathcal{H}^+, \mathcal{H}^-$	NCE	D_{obj}	**0.705**	**0.571**	**0.470**	**0.393**	**0.260**	**0.545**	**1.029**
C	$\mathcal{H}^+, \mathcal{H}^-$	NCE	D'_{obj}	0.701	0.569	0.469	0.392	0.256	0.540	1.023
D	$\mathcal{H}^+, \mathcal{H}^-$	NCE	D_{act}	0.699	0.561	0.453	0.369	0.251	0.538	0.963
E	$\mathcal{H}^+, \mathcal{H}^-$	NCE	D	0.707	**0.571**	0.466	0.385	0.258	0.542	1.020
F	$\mathcal{H}^+, \mathcal{H}^-_{\mathrm{rand}}$	NCE	D_{obj}	0.693	0.557	0.452	0.370	0.253	0.536	0.957
G	$\mathcal{H}^+, \mathcal{H}^0$	NCE	D	**0.705**	0.570	0.466	0.387	0.258	0.542	1.022
H	$\mathcal{H}^+, \mathcal{H}^0_{\mathrm{rand}}$	NCE	D	0.696	0.563	0.462	0.383	0.254	0.536	1.005
I	$\mathcal{H}^+, \mathcal{H}^-$	S-NCE	D_{obj}	0.696	0.561	0.458	0.378	0.252	0.538	0.999
J	$\mathcal{H}^+, \mathcal{H}^-$	L1-PD	D_{act}	0.699	0.569	0.468	0.390	0.255	0.543	1.008

Table 4: **Contrastive learning with counterfactual dialogues:** We experiment with variants of contrastive dialogues pairs, hidden state sampling, and loss. Metrics: B-n: BLEU-n, M: METEOR, R: ROUGE-L, C: CIDEr.

Model	Visual Features	B-1	B-2	B-3	B-4	M	R	C
Baseline (Hori et al., 2019)	I3D	0.621	0.480	0.379	0.305	0.217	0.481	0.733
JMAN (Chu et al., 2020)	I3D	0.648	0.499	0.390	0.309	0.240	0.520	0.890
FA-HRED (Nguyen et al., 2018)	I3D	0.648	0.505	0.399	0.323	0.231	0.510	0.843
Student-Teacher (Hori et al., 2019) †	I3D	0.675	0.543	0.446	0.371	0.248	0.527	0.966
MSTN (Lee et al., 2020) †	I3D	-	-	-	0.379	0.261	0.548	1.028
BiST (Le et al., 2020)	RX	**0.711**	**0.578**	0.475	0.394	0.261	**0.550**	1.050
RLM-GPT2 (Li et al., 2021b) † ‡	I3D	0.694	0.570	**0.476**	**0.402**	0.254	0.544	**1.052**
MTN (Le et al., 2019)	I3D	0.695	0.558	0.455	0.376	0.253	0.534	0.996
MTN $+C^3$ ($\mathcal{I}^{+/-}$)	I3D	0.709	0.577	**0.476**	0.398	**0.262**	0.549	1.040
MTN $+C^3$ ($\mathcal{H}^{+/-}$)	I3D	0.705	0.571	0.470	0.393	0.260	0.545	1.029

Table 5: **Overall results**: † incorporates additional video background audio inputs. ‡ indicates finetuning methods on pretrained language models. Metrics: B-n: BLEU-n, M: METEOR, R: ROUGE-L, C: CIDEr.

nally, to utilize any object-level invariance between $\mathcal{I}^+$ and $\mathcal{I}^-$, we applied a pairwise L1 distance loss $\mathcal{L}^{\mathrm{act}} = \sum_i \|sim(z_i, z_i^+) - sim(z_i, z_i^-)\|_1$ to minimizes distances of hidden states of D_{obj} positions (Row J). However, the performance gain of this loss is not significant, demonstrating representation learning through data variance is a better strategy.

Contrastive Learning with Counterfactual Dialogues. From Table 4, we observed consistent observations as compared to prior experiments with counterfactual videos. Essentially, our results illustrate the impacts of C^3 that specifically contrasts object-level information between $\mathcal{H}^-$ and $\mathcal{H}^+$.

Overall Results. In Table 5, we reported the results of our models which we trained on an MTN backbone (Le et al., 2019) incorporated our proposed C^3 learning approach with counterfactual videos or dialogues. Our models achieve very competitive performance against models trained on the same data features e.g. MSTN (Lee et al., 2020), as well as models pretrained with a large language dataset e.g. RLM-GPT2 (Li et al., 2021b). We also observed that the performance gain of C^3 with $\mathcal{I}^{+/-}$ is higher than that with $\mathcal{H}^{+/-}$. As we showed the benefits of augmented counterfactual dialogues and videos, we will leave the study to unify both augmented data types for a hybrid contrastive learning approach for future work. In this paper, we showed that either dialogues or videos can be augmented and used to improve contextual representations through contrastive losses based on object-based or action-based variance.

For example factual/counterfactual videos/dialogues, please refer to the Appendix.

5 Discussion and Conclusion

In this work, we proposed Compositional Counterfactual Contrastive Learning (C^3), a contrastive learning framework to address the limitation of data in video-grounded dialogue systems. We introduced a factorized object-action structural causal model, described a temporal-based and token-based augmentation process, and formulated contrastive learning losses that exploit object-level and action-level variance between factual and counterfactual training samples. In our proposed approach, we train models to minimize the distance between compositional hidden state representations of factual samples and maximize the distance between counterfactual samples.

We noted our proposed C^3 still entails some limitations. we describe these limitations and suggest potential ways to overcome them for future extension. First, in our approach, we made the assumption of independence between $\mathcal{C}_{obj}$ and $\mathcal{C}_{act}$ to mask tokens/video segments as a way to generate counterfactual data samples. However, in many cases, this assumption might be too strong. Therefore, our approach might disrupt the natural data distribution and create negative noise in model training. A more advanced counterfactual data generation should be able to better capture the nature of counterfactual scenarios, avoiding the above assumption and generalizing the model better. Secondly, in our approach, we require external text-processing tools to decompose the input components. More sophisticated tools could be used to improve data quality of counterfactual/factual examples. Finally, after this work was completed, there have been several more advanced approaches following MTN (Le et al., 2019). As our approach is model-agnostic, we encourage readers to review and adapt our work to these more advanced models.

6 Broader Impacts

In this work, we described C^3, a novel contrastive learning approach that exploits action-based and object-based variance between counterfactual video/dialogue pairs. We demonstrated the benefit of this approach in the video-grounded dialogue domain, which is typically suffered from dataset scarcity. We want to emphasize that our method should be used strictly to improve dataset quality and obtain model performance gains. For instance, a chatbot that incorporates C^3 can generate high-quality responses that better match human questions. Our method should not be used for malicious purposes, such as creating chatbots to steal information or make scam calls.

Considering the widespread application of AI in the real world, the adoption of our method can lead to better dialogue systems that improve the quality of life for many people. For instance, a better chatbot embedded in electronic devices will improve both user experience and productivity. Conversely, the adoption of dialogue systems might lead to the potential loss of jobs in domains such as customer call centres. In high-risk domains such as autonomous vehicles, applications of our method can improve virtual assistant applications in the vehicles. As the products might directly affect human safety, any applications of C^3 should be tested to account for different scenarios, whether the method works as intended or not, and mitigate consequences when the output is incorrect. We advise that any plan to apply our method should consider carefully all potential groups of stakeholders as well as the risk profiles of applied domains to maximize the overall positive impacts.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019a. Audio visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Stefan Lee, Peter Anderson, Irfan Essa, Devi Parikh, Dhruv Batra, Anoop Cherian, Tim K. Marks, and Chiori Hori. 2019b. Audio-visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. 2020. Cophy: Counterfactual learning of physical dynamics. In *International Conference on Learning Representations*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. Evaluating visual conversational agents via cooperative human-ai games. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Yun-Wei Chu, Kuan-Yen Lin, Chao-Chun Hsu, and Lun-Wei Ku. 2020. Multi-step joint-modality attention network for scene-aware dialogue system. *DSTC Workshop @ AAAI*.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28:2296–2304.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 878–892, Online. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. 2020. Contrastive learning for weakly supervised phrase grounding. In *ECCV*.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy. PMLR.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Olivier Henaff. 2020. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

C. Hori, H. Alamri, J. Wang, G. Wichern, T. Hori, A. Cherian, T. K. Marks, V. Cartillier, R. G. Lopes, A. Das, I. Essa, D. Batra, and D. Parikh. 2019. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2352–2356.

Chiori Hori, Anoop Cherian, Tim K Marks, and Takaaki Hori. 2019. Joint student-teacher learning for audio-visual scene-aware dialog. *Proc. Interspeech 2019*, pages 1886–1890.

Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. 2018. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Jiaji Huang, Yi Li, Wei Ping, and Liang Huang. 2018. Large margin neural language model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191, Brussels, Belgium. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron

Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5612–5623, Florence, Italy. Association for Computational Linguistics.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven C.H. Hoi. 2020. BiST: Bi-directional spatio-temporal reasoning for video-grounded dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1846–1859, Online. Association for Computational Linguistics.

Hung Le, Chinnadhurai Sankar, Seungwhan Moon, Ahmad Beirami, Alborz Geramifard, and Satwik Kottur. 2021. Dvd: A diagnostic dataset for multi-step reasoning in video grounded dialogue. *arXiv preprint arXiv:2101.00151*.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. Dstc8-avsd: Multimodal semantic transformer network with retrieval style word generator. *DSTC Workshop @ AAAI 2020*.

Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. 2021. Contrastive learning with adversarial perturbations for conditional text generation. In *International Conference on Learning Representations*.

Shiyang Li, Semih Yavuz, Kazuma Hashimoto, Jia Li, Tong Niu, Nazneen Rajani, Xifeng Yan, Yingbo Zhou, and Caiming Xiong. 2021a. Coco: Controllable counterfactuals for evaluating dialogue state trackers. In *International Conference on Learning Representations*.

Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. 2021b. Bridging text and video: A universal multimodal transformer for video-audio scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–1.

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. 2020. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3285–3292.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Yang Liu and Maosong Sun. 2015. Contrastive unsupervised word alignment with non-local features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2020. Generate your counterfactuals: Towards controlled counterfactual generation for text. *arXiv preprint arXiv:2012.04698*.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27:1682–1690.

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Dat Tien Nguyen, Shikhar Sharma, Hannes Schulz, and Layla El Asri. 2018. From film to video: Multi-turn question answering with multi-modal context. In *AAAI 2019 Dialog System Technology Challenge (DSTC7) Workshop*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press, USA.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5043–5053, Hong Kong, China. Association for Computational Linguistics.

Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. 2019. Factor graph attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2039–2048.

Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. 2017. Visual reference resolution using attention memory for visual dialog. In *Advances in neural information processing systems*, pages 3719–3729.

Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1049–1058.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Zhao Wang and Aron Culotta. 2020. Robustness to spurious correlations in text classification via automatically generated counterfactuals. *arXiv preprint arXiv:2012.10040*.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Florence, Italy. Association for Computational Linguistics.

Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S Lasecki, Jonathan K Kummerfeld, Michel Galley, Chris Brockett, et al. 2019. Dialog system technology challenge 7. *arXiv preprint arXiv:1901.03461*.

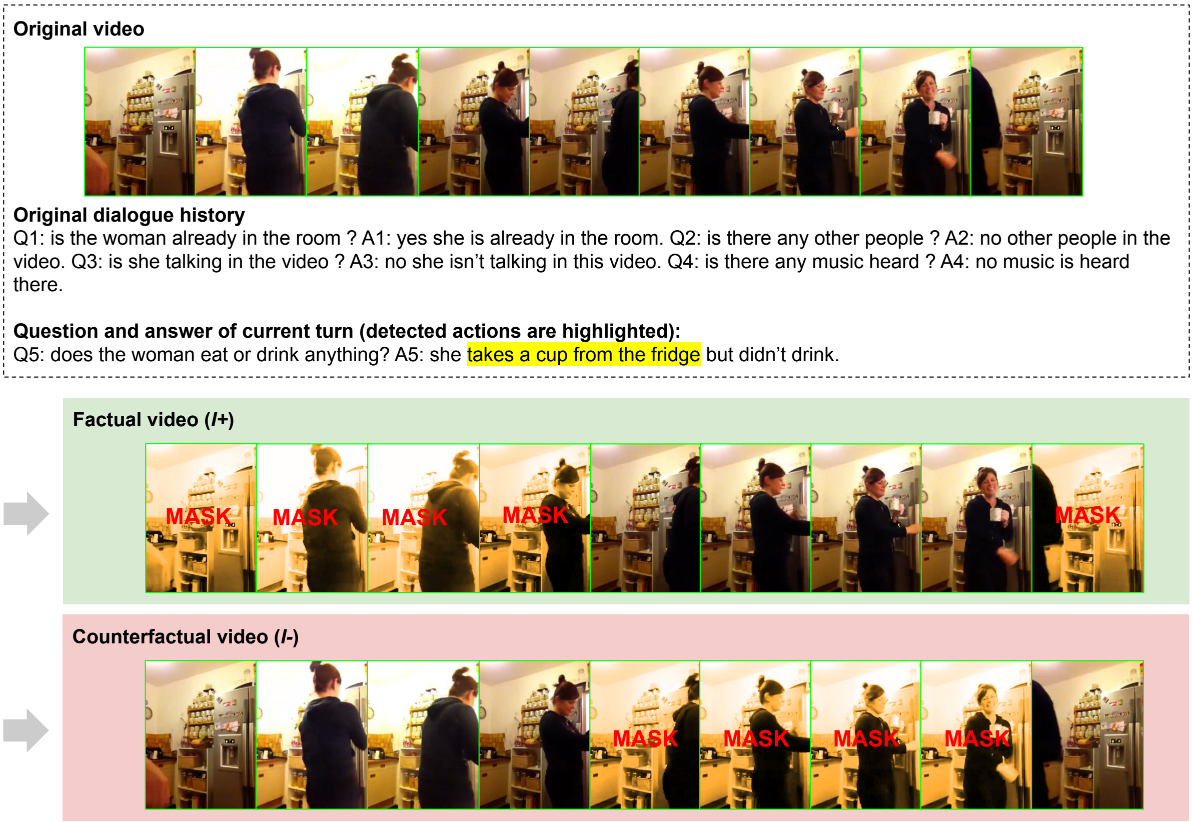
Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280, Online. Association for Computational Linguistics.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022.

Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. 2020. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems*, 33:18123–18134.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.



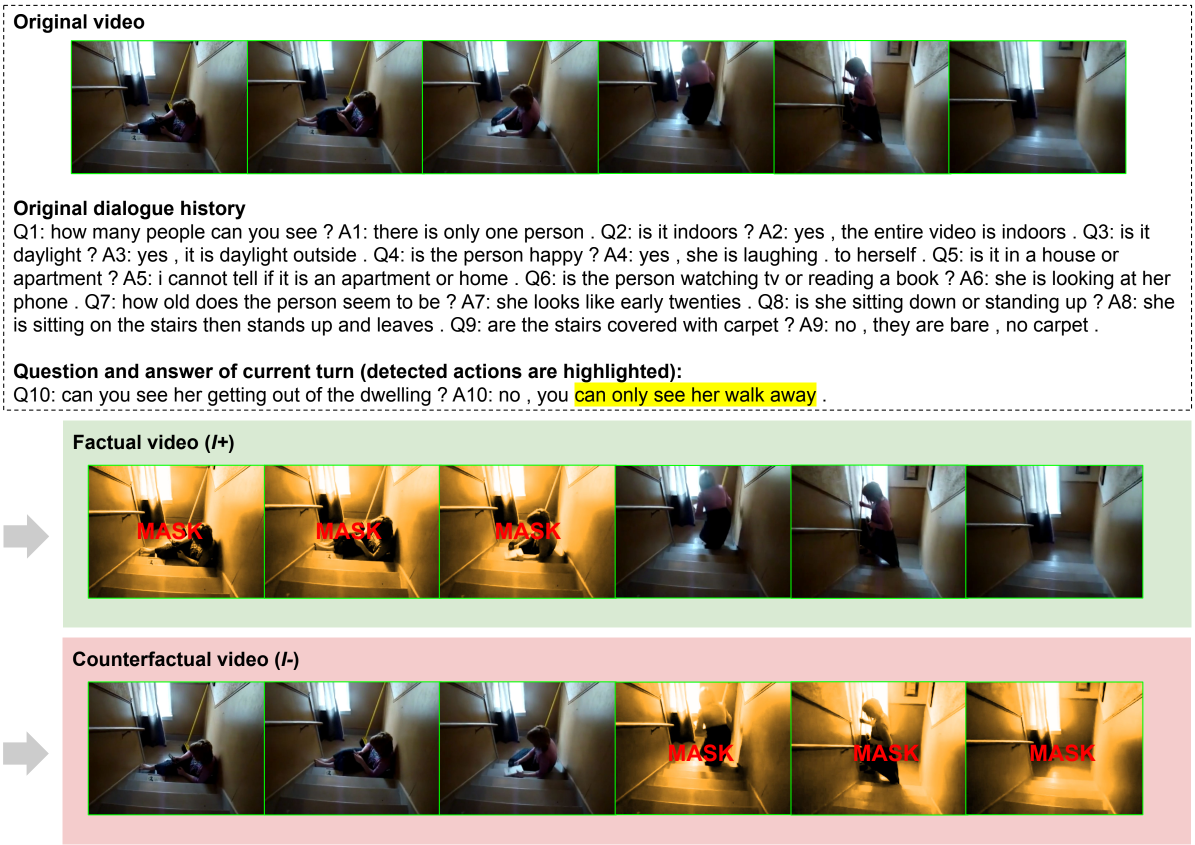
Original dialogue history

Q1: is the woman already in the room ? A1: yes she is already in the room. Q2: is there any other people ? A2: no other people in the video. Q3: is she talking in the video ? A3: no she isn't talking in this video. Q4: is there any music heard ? A4: no music is heard there.

Question and answer of current turn (detected actions are highlighted):

Q5: does the woman eat or drink anything? A5: she takes a cup from the fridge but didn't drink.

Figure 4: Example factual and counterfactual video



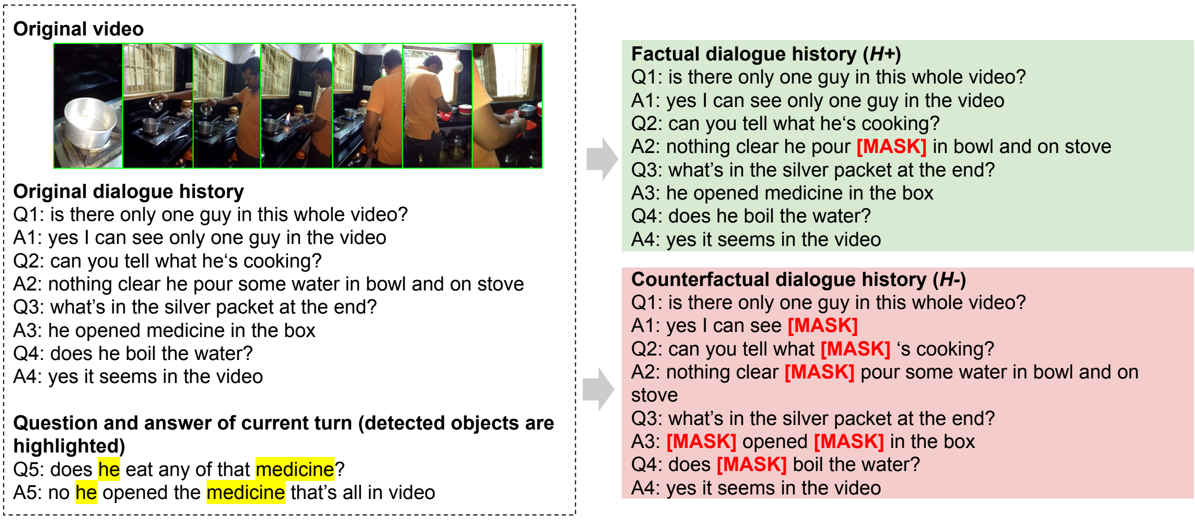
Original dialogue history

Q1: how many people can you see ? A1: there is only one person . Q2: is it indoors ? A2: yes , the entire video is indoors . Q3: is it daylight ? A3: yes , it is daylight outside . Q4: is the person happy ? A4: yes , she is laughing . to herself . Q5: is it in a house or apartment ? A5: i cannot tell if it is an apartment or home . Q6: is the person watching tv or reading a book ? A6: she is looking at her phone . Q7: how old does the person seem to be ? A7: she looks like early twenties . Q8: is she sitting down or standing up ? A8: she is sitting on the stairs then stands up and leaves . Q9: are the stairs covered with carpet ? A9: no , they are bare , no carpet .

Question and answer of current turn (detected actions are highlighted):

Q10: can you see her getting out of the dwelling ? A10: no , you can only see her walk away .

Figure 5: Example factual and counterfactual video



Original video

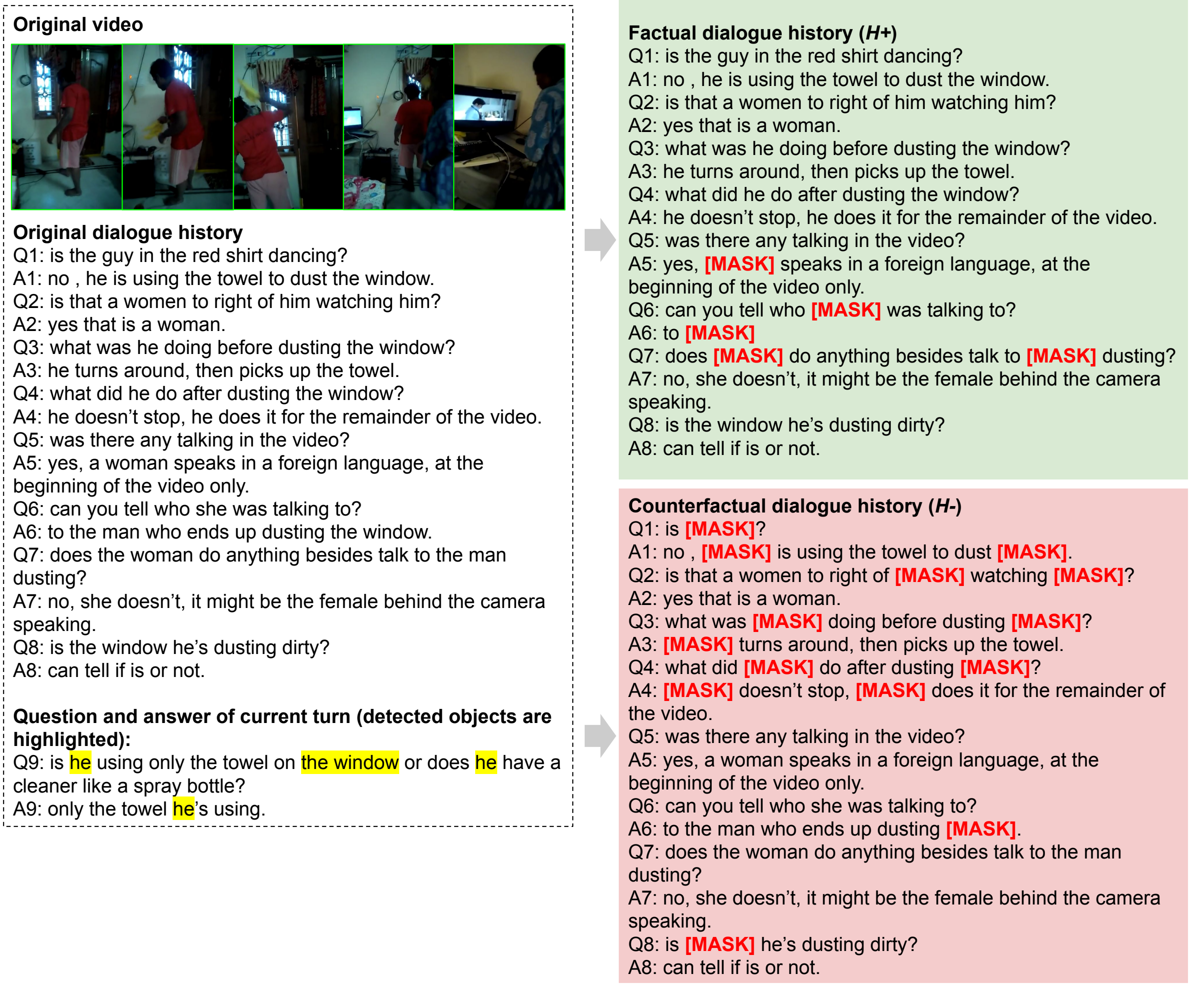
Original dialogue history
Q1: is there only one guy in this whole video?
A1: yes I can see only one guy in the video
Q2: can you tell what he's cooking?
A2: nothing clear he pour some water in bowl and on stove
Q3: what's in the silver packet at the end?
A3: he opened medicine in the box
Q4: does he boil the water?
A4: yes it seems in the video

Question and answer of current turn (detected objects are highlighted)
Q5: does he eat any of that medicine?
A5: no he opened the medicine that's all in video

Factual dialogue history (*H+*)
Q1: is there only one guy in this whole video?
A1: yes I can see only one guy in the video
Q2: can you tell what he's cooking?
A2: nothing clear he pour **[MASK]** in bowl and on stove
Q3: what's in the silver packet at the end?
A3: he opened medicine in the box
Q4: does he boil the water?
A4: yes it seems in the video

Counterfactual dialogue history (*H-*)
Q1: is there only one guy in this whole video?
A1: yes I can see **[MASK]**
Q2: can you tell what **[MASK]** 's cooking?
A2: nothing clear **[MASK]** pour some water in bowl and on stove
Q3: what's in the silver packet at the end?
A3: **[MASK]** opened **[MASK]** in the box
Q4: does **[MASK]** boil the water?
A4: yes it seems in the video

Figure 6: Example factual and counterfactual dialogue history



Original video

Original dialogue history
Q1: is the guy in the red shirt dancing?
A1: no , he is using the towel to dust the window.
Q2: is that a women to right of him watching him?
A2: yes that is a woman.
Q3: what was he doing before dusting the window?
A3: he turns around, then picks up the towel.
Q4: what did he do after dusting the window?
A4: he doesn't stop, he does it for the remainder of the video.
Q5: was there any talking in the video?
A5: yes, a woman speaks in a foreign language, at the beginning of the video only.
Q6: can you tell who she was talking to?
A6: to the man who ends up dusting the window.
Q7: does the woman do anything besides talk to the man dusting?
A7: no, she doesn't, it might be the female behind the camera speaking.
Q8: is the window he's dusting dirty?
A8: can tell if is or not.

Question and answer of current turn (detected objects are highlighted):
Q9: is he using only the towel on the window or does he have a cleaner like a spray bottle?
A9: only the towel he's using.

Factual dialogue history (*H+*)
Q1: is the guy in the red shirt dancing?
A1: no , he is using the towel to dust the window.
Q2: is that a women to right of him watching him?
A2: yes that is a woman.
Q3: what was he doing before dusting the window?
A3: he turns around, then picks up the towel.
Q4: what did he do after dusting the window?
A4: he doesn't stop, he does it for the remainder of the video.
Q5: was there any talking in the video?
A5: yes, **[MASK]** speaks in a foreign language, at the beginning of the video only.
Q6: can you tell who **[MASK]** was talking to?
A6: to **[MASK]**
Q7: does **[MASK]** do anything besides talk to **[MASK]** dusting?
A7: no, she doesn't, it might be the female behind the camera speaking.
Q8: is the window he's dusting dirty?
A8: can tell if is or not.

Counterfactual dialogue history (*H-*)
Q1: is **[MASK]**?
A1: no , **[MASK]** is using the towel to dust **[MASK]**.
Q2: is that a women to right of **[MASK]** watching **[MASK]**?
A2: yes that is a woman.
Q3: what was **[MASK]** doing before dusting **[MASK]**?
A3: **[MASK]** turns around, then picks up the towel.
Q4: what did **[MASK]** do after dusting **[MASK]**?
A4: **[MASK]** doesn't stop, **[MASK]** does it for the remainder of the video.
Q5: was there any talking in the video?
A5: yes, a woman speaks in a foreign language, at the beginning of the video only.
Q6: can you tell who she was talking to?
A6: to the man who ends up dusting **[MASK]**.
Q7: does the woman do anything besides talk to the man dusting?
A7: no, she doesn't, it might be the female behind the camera speaking.
Q8: is **[MASK]** he's dusting dirty?
A8: can tell if is or not.

Figure 7: Example factual and counterfactual dialogue history